Copyright Infringement Determination in Text Content Generated by Artificial Intelligence

Yilong Song

The Law School of Heilongjiang University, Harbin 150080, China

yilong-song@qq.com

Abstract. Copyright issues of artificial intelligence-generated content have become a key focus in AI-related rule of law. Of these, copyright violations of AI-generated text are especially pertinent. The old test, of “access + substantial similarity” cannot be correctly applied to the text generated by AI because of the nature of such text. The U.S. system is based on case law; the EU and Germany systems are based on comprehensive institutional frameworks and have different approaches. In this way, infringing on such a basis can be comprehensively determined by distinguishing protected and unprotected elements of copyrighted works and identifying original elements. A filing and inspection mechanism for AI training data should also be put in place to ease the determination of the "access" requirement.

Keywords: Artificial Intelligence Generated Content, Substantial Similarity, Copyright Infringement.

1. Introduction

With advances in science and technology, artificial intelligence has grown explosively in recent years. Since 2016, the match between AlphaGo and top chess players has drawn wide public attention, bringing AI into public view. After OpenAI launched ChatGPT in 2022, generative artificial intelligence (AI) has swept across the globe rapidly. Internet giants such as Google, Tencent, and Alibaba have successively released their own large AI models. These models generally adopt a question-and-answer interaction mode, providing a more natural human-computer experience and driving broader changes in industrial production, content creation, and other fields.

In content creation, AI-generated content (AIGC) covers videos, images, audio, text, and other forms, but text remains the dominant type of AI creation. AIGC text includes news reports, essays, poems, copywriting, and other genres. It is widely available and easy to use, so it has greatly impacted traditional text creation and triggered a series of legal issues. Most of these issues relate to copyright, including whether AI-generated text qualifies as a work, how to determine copyright infringement, and how to protect copyright. Since current large AI models mainly use question-and-answer modes and focus on text output, determining whether AIGC text infringes existing copyrights has become central to AI copyright issues.

Internationally, research on AI copyright has a long history, and AI infringement has become a research hotspot in recent years. In the early stage of AI development, the United States established the Committee on New Technologies Uses of Copyrighted Works (CONTU) in 1974 to study impacts of new technologies, especially computer technology, on the copyright system. However, in its 1979 report to Congress, CONTU held that copyright law did not need special provisions for machine-generated works[1]. In 1981, American scholar Buffer discussed copyright of AI-generated content in *Can a Computer Be an Author?*[2] The United Kingdom stipulated in Section 9 of Chapter 48 of the Copyright, Designs and Patents Act 1988 that copyright in computer-generated literary and artistic works belongs to the person who made the necessary arrangements for its creation. It was the first country to clearly grant copyright to computer-generated content[3].

Since the 21st century, improved computing power has encouraged governments and organizations to build AI copyright systems. In 2016, Japan's Intellectual Property Strategy Headquarters tried to establish a registration system for AI-generated content to replace relevant regulations in the Copyright Act[4]. In 2019, the International Association for the Protection of Intellectual Property confirmed in its *Resolution on Copyright Issues of AI-Generated Content* that AI-generated content

is protected by copyright if it involves human intervention and originality derived from such intervention[5]. On copyright infringement of AIGC, the EU listed copyright compliance as an access condition for high-risk AI systems in the AI Act 2024, providing a reference for regulating AIGC copyright infringement. Overall, the international community has not formed a unified and complete theoretical system for AI copyright, and differences remain in various aspects.

In China, research on AI copyright emerged around 2017, starting relatively late and focusing mainly on theory. Studies on copyright protection and infringement of AIGC center on three issues: whether to grant copyright protection to AI-generated content, how to protect it, and who owns the copyright[6]. In judicial practice, relevant cases have emerged, such as the AI text-to-image case heard by Beijing Internet Court in 2023. However, there is still no specific and feasible standard for determining and regulating AIGC copyright infringement, leading to inconsistent judicial practices. Compared with international research, Chinese studies remain largely theoretical, with limited legislation and judicial rules.

China is now in a period of rapid AI development. Although academic circles have explored AI copyright, most research is theoretical and insufficient regarding AIGC text as a specific object, requiring in-depth study. This paper focuses on AIGC text. Based on typical cases of AIGC text copyright infringement, it discusses how to determine infringement at the content output stage. It aims to define the infringement boundary of AIGC text copyright on the basis of existing research.

2. Judicial Determination of AIGC Text Copyright Infringement in China

2.1. Analysis of Determination Logic in AIGC Copyright Infringement Cases

So far, China has published few AI copyright infringement cases, and there is no effective judgment directly involving AIGC text infringing others' copyright. Existing cases fall into two categories: one involves infringement of copyright in AI-generated content by others, with the infringement direction opposite to that discussed in this paper; the other involves AI-generated content infringing others' copyright, but the objects are mostly images or videos rather than text. Nevertheless, analysis of decided AI-related cases can summarize the basic logic followed by Chinese courts in handling AIGC infringement cases, providing a reference for future AIGC text copyright infringement cases.

The case of Shanghai Xinchuanghua Culture Development Co., Ltd. v. Guangzhou Nianguang Company[7] (hereinafter the Xinchuanghua Case) is representative in image-based AIGC copyright infringement. It reflects the court's approach to AIGC copyright infringement. The court first confirmed ownership of the Ultraman image. It then contrasted the image made by Tab with the copyrighted Ultraston art work and determined there was a high degree of similarity. Based on the fame of Ultraman works, the court found the "access" requirement satisfied. It found that the defendant had violated the plaintiff's right of reproduction and since the AI device had generated new expression, also the right of adaptation. Lastly, the court considered the party liable and apportioned liability. It found that the defendant's conduct did not meet the standard of reasonable duty of care, and was subjectively at fault and ordered it to cease infringement and pay damages, based on the Interim Measures for the Administration of Generative Artificial Intelligence Services. This case, while dealing with images, has a legal reasoning that is relevant to text-based AIGC copyright infringement.

There have been no direct cases of copyright violations to AIGC text. But, in related cases, court reasoning sheds some light on originality. Tencent filed a lawsuit against Yingxun for unauthorized printing of financial articles created by Tencent's Dreamwriter, claiming it infringed its copyright in the Dreamwriter Case [8]. Ownership and alleged infringement was first confirmed by court. The court did not discuss whether it was an AI-generated article or not, unlike the Xinchuanghua Case. It's originality is telling. The court explored the concepts of independent creation and minimum creativity, and then considered whether the work was the work of the creator's own choice, judgment and skill. It then held that the article, "AI-generated or not, is considered original and is a work.

In Beijing Feilin Law Firm v. Beijing Baidu Netcom Technology Co. , Ltd. [9] (hereinafter the Weike Xianxing Case), the court once again examined the question of originality of the work created by the Weike Xianxing database, and the court upheld originality. This demonstrates that the basic standard for originality in China is personalized choice, judgment and skill, and is thus equally applicable to substantial similarity.

In summary, Chinese courts follow three steps in AIGC copyright infringement: first, confirm ownership of the copyrighted work to establish procedural standing of the plaintiff and defendant; second, review whether infringement exists, mainly judging substantial similarity and access; third, examine the defendant's fault and determine liability form and damages. This framework—right basis review → infringement constitution review → liability review and allocation—provides a basic structure for determining AIGC text copyright infringement.

2.2. Core Elements of AIGC Text Copyright Infringement Determination

Copyright infringement generally requires four elements: fault, infringing act, harmful result, and causation. The elements in the above three cases are shown in Table 1.

Table 1. Analysis of Infringement Elements in AIGC Copyright Cases

Case	AI Tool	Fault	Unlawful Act	Harmful Result	Causation	Content Form	Right Owner
Xinchuanghua Case	Tab	Yes	Accessed and partially or fully copied the original expression of the Ultraman image	Infringed right of reproduction and adaptation	Yes	Image	Plaintiff
Dreamwriter Case	Dream writer	No	Published the article online without authorization	Infringed right of information network transmission	Yes	Text	Plaintiff
Weike Xianxing Case	Weike Xianxing	No	Provided the article on its Baijiahao platform without authorization	Infringed right of information network transmission, right of authorship, and right of integrity	Yes	Text	Plaintiff

On this basis, core elements for determining AIGC text copyright infringement can be extracted. Compared with the above cases, AIGC text infringement has both common and unique features.

The "access + substantial similarity" test still applies to determining infringing acts and originality in AIGC text. Substantial similarity is more complex for literary works than for images. In the Xinchuanghua Case, the court compared main features to find similarity. For literary works, courts must use the idea-expression dichotomy and identify original expression. Specifically, if AIGC text completely copies original expression of existing works (similar to the Dreamwriter Case), it may infringe the right of reproduction. If AI interactively provides copyrighted works without authorization, it may infringe the right of information network transmission. For access, courts can presume access based on the fame of the original work in AIGC text cases. Access is easily presumed for text works published online, but plaintiffs may bear the burden of proof for unpublished works. In summary, Chinese courts still apply the "access + substantial similarity" test in AIGC infringement.

2.3. Difficulties in Determining AIGC Text Copyright Infringement

Although some cases provide reference, the traditional "access + substantial similarity" rule faces new challenges in identifying AIGC text infringement and originality.

For the "access" requirement, traditional courts verify whether the defendant had access to the plaintiff's work to judge originality. In AIGC, users and service providers may not access prior works; instead, AI developers input prior works during training. Courts can presume access based on fame (as in the Xinchuanghua Case), but how to judge access if the text is not famous or unpublished? In practice, the "access" requirement is often undermined by "prior publication": courts tend to presume access if the work was published online [10]. This tendency is stronger in AIGC text cases because AI training draws massive works from the Internet, and online AI services directly use others' works,

making it hard to rule out prior "access". However, if AI generates content by analyzing existing texts or creates new content through "learning", this logic may lead to misjudgment of "access". Such misjudgment mainly stems from insufficient disclosure of AI training data sources, leaving no direct evidence for "access" and forcing reliance on indirect inference from prior publication.

For "substantial similarity", traditional methods are based on human creation. AIGC text may retrieve and reorganize content to infringe multiple copyright owners in one text, with the infringing portion being small in each original work[11]. AI may also imitate the overall structure of a work while greatly changing specific details, increasing difficulty of similarity judgment. Moreover, AI training does not distinguish between copyrighted and public-domain works, so generated content may mix protected and unprotected parts. Existence of large unprotected portions may interfere with identification of protected content in substantial similarity review.

3. International Determination and Practice of AIGC Text Copyright Infringement

With global development of AI, AIGC copyright disputes have emerged worldwide. As major economies and pioneers in AI, the United States and the EU have rich legislation and judicial practice in AIGC text copyright infringement. This paper takes the US and the EU as comparative objects to extract rules for China and solve domestic difficulties in AIGC text infringement determination.

3.1. Determination in the United States

As a global hub for AI research and development, the US has not enacted special AIGC legislation. However, substantial similarity standards formed in case law are flexible and suitable for resolving copyright issues brought by rapid AI progress.

A preliminary ruling in the 2025 *Authors Guild v. OpenAI* case directly shows how US courts apply existing rules to AIGC text copyright infringement (the ruling is effective, but the case is still in expert discovery). The Authors Guild sued OpenAI on behalf of writers for using copyrighted literary works without permission and generating infringing content. In October 2025, the US District Court for the Southern District of New York issued a preliminary ruling. It held that the class complaint, based on ChatGPT output, reasonably alleged prima facie copyright infringement by OpenAI and dismissed OpenAI's objections[12]. The court explained two tests for substantial similarity: the "ordinary observer test" and the "more discerning observer test". It held that when a work contains both protected and unprotected elements, the "more discerning observer test" should be used to compare original protected elements for substantial similarity. The court analyzed threshold and substantive issues and confirmed that the more discerning observer test applies because the plaintiffs' works include both protected and unprotected elements.

In this case, summaries generated by ChatGPT did not copy every plot or character element of the original works, but they abridged or condensed core protected content. A more discerning observer could easily find substantial similarity between the ChatGPT summaries and Martin's original works. The outline of a potential sequel generated by the AI was also substantially similar. The court thus held that a reasonable jury applying the more discerning observer test could find substantial similarity, as the AI output incorporated protected elements of Martin's works.

Originating from US case law, the more discerning observer test separates protected and unprotected elements for complex copyrighted works, reducing identification difficulty. It then compares the overall impression of core protected elements from the perspective of an observer capable of recognizing originality. This helps resolve substantial similarity disputes when AIGC uses parts of original works.

3.2. Determination in the EU

The EU is a leader in AI legislation and has built a multi-level regulatory system for AI copyright. Based on the *InfoSoc Copyright Directive* and the *DSM Copyright Directive*, supplemented by the *AI*

Act 2024, the system clearly defines copyright protection in the digital era. Article 2 of the InfoSoc Directive expands the right of reproduction to any form or manner, adapting to digital copying. Article 53(1)(d) of the AI Act requires AI service providers to document and disclose training content, helping courts identify and protect input works used in AI training.

As a core EU member, Germany transposes EU directives into domestic law. In 2025, it ruled in *GEMA v. OpenAI* that the operator of an AI model is liable for copyright infringement. Although this is a first-instance judgment, its reasoning is instructive. GEMA sued OpenAI for using nine copyrighted lyrics to train AI and generate related content without permission. On reproduction infringement, the court clearly held in paragraph 176 that the AI-generated lyrics constituted reproduction under Article 16 of the German Copyright Act, which implements Article 2 of the EU InfoSoc Directive. The court uniformly interpreted the EU concept of reproduction to include any manner and temporary or permanent copying.

In this case, lyrics used as training data were reproducibly fixed in the AI model, falling within "any form" and satisfying the "reproducible fixation" requirement for reproduction. Use of the model materialized lyrics as output, allowing human perception of the fixed text. The court thus held that use of the model constituted reproduction of the original lyrics. For similarity between AIGC and original lyrics, the court held that the key is whether creative elements of the protected work are recognizably used in the accused object. After line-by-line comparison, the court found that AI output mostly reproduced original lyrics verbatim. Additional content and punctuation changes did not obscure originality, so the output was not independent creation but copying.

Germany does not use the "access + substantial similarity" test. Instead, it judges similarity by whether original elements of the protected work can be identified in the accused content. This focuses on original details and avoids errors from overall impression. It holds that extra AI output does not prevent identification of original elements, thus determining reproduction.

In summary, US courts use the ordinary observer test and more discerning observer test for substantial similarity. The latter separates protected elements before overall review, helping when AIGC uses parts of works. The EU has more complete AI legislation, gradually implemented by member states. The InfoSoc Directive expands reproduction rights to cover digital changes. German courts apply laws transposed from EU directives. They judge similarity by identifying original elements in accused content, reducing interference from extra AI output.

4. Improvement of China's AIGC Text Copyright Infringement Standards

Facing evolving AI technology and increasingly complex applications, China's current copyright system urgently needs an institutionalized and systematic legal framework to address difficulties in determining and regulating AIGC text copyright infringement. The core of AIGC text infringement determination is to improve the "access + substantial similarity" test so that it adapts to new infringement issues and facilitates judicial judgment. Systems of the US, EU, and Germany provide valuable experience.

4.1. Optimizing the Substantial Similarity Test

As the global center of research and development for the artificial intelligence industry, the United States has not enacted legislation specifically addressing AIGC. Nevertheless, the substantial similarity standards developed through its case law are more flexible and are therefore better suited to resolving copyright infringement issues arising from the rapid advancement of AI technologies[13].

A preliminary ruling in Authors Guild v. OpenAI in 2025 directly illustrates how U.S. courts employ existing legal rules to address copyright infringement issues involving AIGC-generated texts. This ruling has already taken effect, although the case remains at the expert discovery stage. In this case, the Authors Guild, on behalf of multiple authors, sued OpenAI, alleging that it had used protected literary works without authorization and generated infringing content. In October 2025, the U.S. District Court for the Southern District of New York issued a preliminary ruling, holding that,

based on content generated by ChatGPT, the consolidated complaint plausibly stated a prima facie claim of copyright infringement against OpenAI and rejecting OpenAI's related objections.

In particular, the court explained two tests for determining substantial similarity: the "ordinary observer test" and the "more discerning observer" test. The latter refers to a method applied where a work contains both protectable and unprotectable elements; by comparing the overall concept, feel, theme, characters, plot, sequence, and pace of the works, the court determines whether the protectable elements of the allegedly infringing work are substantially similar to those of the original work. The court held that when a work contains both protectable and unprotectable elements, the "more discerning observer" test should be applied to compare those original elements that are entitled to copyright protection and to determine whether substantial similarity exists. It then analyzed both the threshold issues and the substantive elements of the claim, concluding that because the plaintiffs' allegedly infringed works contained both copyright-protected and unprotected elements, the "more discerning observer" test was the appropriate standard.

IThe United States is the hub of artificial intelligence industry research and development, and there is no specific legislation on AIGC. However, its case law has established substantial similarity standards which are more flexible, and well suited to solve copyright infringement problems in relation to the fast development of AI technologies[14].

Author's Guild v. OpenAI (2025) is a preliminary decision by a U. S. court that directly relates to copyright infringement concerns with AIGC-generated texts. This decision has already gone into effect, but the case is still at the expert discovery phase. In this instance, the Authors Guild, representing several authors, filed a lawsuit against OpenAI, claiming that it had copied works of literature and created infringing content without the authors' consent. The U. S. District Court for the Southern District of New York issued a preliminary ruling in October 2025 finding that, based on content created by ChatGPT, the consolidated complaint stated a prima facie case of copyright infringement against OpenAI and dismissed OpenAI's objections to the case.

In particular, the court explained two tests for determining substantial similarity: the "ordinary observer test" and the "more discerning observer" test. The latter is a method that is applied when a work consists of both protectable and unprotectable elements, where the court compares the overall concept, feel, theme, characters, plot, sequence and pace of the works to see if the protectable elements of the allegedly infringing work are substantially similar to the original work. The court found that in the case of works with protectable and unprotectable components, the "more discerning observer" test should be used to compare the original elements that are subject to copyright protection, and to determine substantial similarity. It then considered the threshold questions as well as the elements of Plaintiffs' allegedly infringed works and concluded that since the allegedly infringed works had both protected and unprotected elements, the "more discerning observer" test was the standard to be applied.

In this instance ChatGPT's summaries did not copy all parts of the plot or character development in the original texts, but instead were actually shortened or abridged versions of the core protected content in the works. A more astute observer would easily draw a conclusion that the summaries written by ChatGPT are significantly alike Martin's original works, and that the outlines of possible sequels written by ChatGPT are similarly significantly alike the originals. Accordingly, the court reasoned that because the AI-generated content incorporated copyright-protected elements from Martin's works, a jury applying the more discerning observer standard could reasonably find substantial similarity between the generated content and Martin's original works[14].

Although the more discerning observer test originated in U.S. case law, it reduces the difficulty of assessment in cases involving complex copyrighted works by separating protectable from unprotectable elements where multiple kinds of elements are intertwined. It then proceeds from the perspective of an observer capable of identifying a work's originality to compare the overall impression created by the core protected elements. This approach helps address the problem of determining substantial similarity where AIGC partially draws upon preexisting works.

4.2. Promoting Enterprises to Establish an AI Training Data Filing and Inspection Mechanism

As noted, Article 53 of the EU AI Act requires disclosure of AI training data sources and mandates publication of training content. Recital 107 requires AI model providers to prepare and publish detailed summaries of training content without compromising trade secrets, so that legitimate stakeholders can exercise rights. This mechanism has a basis in China's current rules. Article 19 of the Interim Measures requires AI service providers to explain training data sources and provide technical assistance. Article 5.2(b)(6) of the Basic Requirements for Generative AI Service Security encourages disclosure of summary information on copyrighted corpus. These provisions impose a duty to explain training data sources and support disclosure, but they are incomplete. When infringement occurs, authorities cannot accurately trace training data sources for infringing AIGC text, and even AI companies cannot trace sources, making the disclosure duty ineffective.

This paper advocates that AI enterprises establish an AI training data filing and inspection mechanism. Enterprises should build supporting databases to record and archive use of copyrighted works in real time during AI training. When AIGC involves infringement litigation, the court may query the training data. Some scholars suggest that copyright owners should apply for access, but this risks data leakage or malicious inquiry. Court inquiry at the access-determination stage is more appropriate[15]. If the work was not used in training, access during training is ruled out; if it was used, access is directly confirmed, simplifying judgment and avoiding misdecision. The Ministry of Industry and Information Technology may conduct random inspections to verify consistency between filed records and actual training data, preventing false filing that undermines judicial judgment. If an enterprise files falsely, its records may not be used for "access" determination. If the original work was publicly available, the enterprise then bears the burden of proving non-access, increasing difficulty and encouraging truthful filing.

This disclosure mechanism serves three purposes: It facilitates legal accountability by verifying whether the accused AI system copied the copyrighted work, providing direct evidence of "access"[16]. It helps enterprises manage training data, review and correct data, and improve technology. Courts only query training data ex officio when judging infringement, reducing leakage risk and protecting trade secrets.

5. Conclusion

AI technology is one of the most dynamic production factors in recent years and plays an irreplaceable role in advancing human production. However, it also brings unavoidable mass copyright infringement. The root cause is misalignment between AI technology and the current copyright system. This not only triggers numerous copyright disputes but also hinders AI development. Starting from solving AIGC text infringement, this paper combines China's current system and judicial reasoning with legislative and judicial experience of the US, EU, and Germany. It improves China's copyright infringement test to adapt to new issues of AIGC text, providing a new approach for infringement determination.

Optimizing the substantial similarity test and establishing an AI training data filing and inspection mechanism can effectively resolve current AIGC text copyright infringement problems in China. They balance AI development and copyright protection, reduce barriers from the current copyright system, promote the development of China's AI industry, and provide legal guarantee for AI progress.

References

[1] Wu Handong, Questions Concerning Copyright Law on Artificially Generated Works, Peking University Law Journal, Issue 3, 2020, p. 664. Quoted from National Commission on New Technological Uses of Copyrighted Works (CONTU), Final Report on New Technological Uses of Copyrighted Works (1979), p. 44.

[2] Timothy L. Butler, Can a Computer be an Author—Copyright Aspects of Artificial Intelligence, 4 Journal of Communications and Entertainment Law 707 (1981), p. 735.

[3] Copyright, Designs and Patents Act 1988, Chapter 48, Section 9(3): In the case of a literary, dramatic, musical or artistic work which is computer-generated, the author shall be taken to be the person by whom the arrangements necessary for the creation of the work are undertaken.

[4] Yu Fenglei, Legislative Research on Copyright Protection of AI-Generated Content in China and Japan, Science, Technology and Law, Issue 1, 2020, p. 4.

[5] APPI World Congress, Resolution 2019: Study Question on Copyright in Artificially Generated Works (Sept. 18, 2019).

[6] Shi Weijia, Research on Copyright Protection of Text Content Generated by Artificial Intelligence [D], Tianjin University, 2021, p. 5.

[7] Guangzhou Intellectual Property Court, Civil Judgment, Case No. (2024) Yue 0192 Min Chu No. 113 [EB/OL]. (2024-05-27) [Visited: 2026-02-16].

[8] Nanshan District People's Court of Shenzhen City, Guangdong Province, Civil Judgment, Case No. (2019) Yue 0305 Min Chu No. 14010 [EB/OL]. (2019-12-24) [Visited: 2026-02-17].

[9] Beijing Intellectual Property Court, Civil Judgment, Case No. (2019) Jing 73 Min Zhong No. 2030 [EB/OL]. (2020-08-06) [Visited: 2026-02-20].

[10] Liu Lin, Review and Reconstruction of the "Access" Element in Copyright Infringement in China [J], Intellectual Property, Issue 11, 2021, p. 74.

[11] Shao Honghong, Research on Copyright Infringement Governance of Generative Artificial Intelligence [J/OL], Publishing & Distribution Research, Issue 6, 2023, pp. 29-38, p. 30.

[12] Authors Guild et al. v. OpenAI Inc. et al., 25-md-3143 (SHS) (OTW) [EB/OL]. (2025-10-27) [Visited: 2026-03-20].

[13] Yao Ye, Is Machine Piracy a Noble Offense?—Identification of Copyright Infringement of AI-Generated Content and Risk Mitigation [J], Jinan Journal (Philosophy & Social Sciences Edition), Vol. 47, Issue 8, 2025, p. 87.

[14] Wang Yuan, Research on the Evolution of U.S. Copyright Infringement Identification Standards [J/OL], Publishing Science, Vol. 24, Issue 1, 2016, p. 28.

[15] Guo Yuxin & Long Mingxia, Copyright Infringement Liability for Mass Use of Works by Generative AI: Trends, Challenges and Countermeasures [J/OL], Publishing & Distribution Research, Issue 10, 2025, p. 21.

[16] Li An. Disclosure of Copyright Information of Artificial Intelligence Training Data: Theoretical Basis and Institutional Arrangements [J]. Journal of Comparative Law Studies, 2024(5): 143.